Selected Issues in Multivariate Analyses That Pertain to Multiple Regression Analysis: Specification and Measurement Errors and Collinearity and Suggestions Concerning How to Deal With Them

Kenneth H. Strand

Illinois State University

Paper presented at the annual meeting of the

American Educational Research Association

New Orleans, April 2000

Abstract

This paper contains information concerning the following:

1. An overview of multivariate analysis of variance, and discriminant (DA) and canonical (CA) analyses.

2. An introduction to specification and measurement errors, and collinearity.

3. The sparsity of information concerning specification and measurement errors and collinearity as they pertain to DA and CA.

4. Selected suggestions regarding how to deal with specification and measurement errors and collinearity as they pertain to DA and CA.

Selected Issues in Multivariate Analyses That Pertain to Multiple Regression Analysis: Specification and Measurement Errors and Collinearity and Suggestions Concerning How to Deal with Them

Considerable recent literature exists regarding (a) the use of multivariate analysis of variance (MANOVA) and discriminant (DA) and canonical (CA) analyses in actual practice, and (b) details concerning the methods per se. A sampling regarding the latter are the works of Chateau (1999); Strand (1999); Humphries-Wadsworth (1998); Fung and Gu (1998); Thompson (1991); Ziari, Leatham, and Ellinger (1997); Pedhazur (1997); Seo, Kanda, and Fujikoshi (1995); Cole, Maxwell, Arvey, and Salas (1994); and Huberty and Wisenbaker (1992).

This paper contains information concerning the following:

1. An overview of MANOVA, DA, and CA.
2. An introduction to specification errors, including measurement errors and collinearity.
3. The sparsity of information concerning specification errors as they pertain to DA and CA.
4. Selected suggestions regarding how to deal with specification errors as they pertain to DA and CA.

Overview of MANOVA, DA, and CA

MANOVA, DA, and CA are multivariate statistical techniques that are related to one another. MANOVA generally pertains to the relationship between one or more categorical independent (X) variables, and multiple continuous dependent (Y) variables.

DA generally pertains to the relationship between a single categorical Y variable and multiple continuous X variables. However, many users feel comfortable about the use of

categorical X variables as long as they are coded appropriately. Furthermore, one may look upon a discriminant analysis as a "flip side" of a one-way MANOVA -- the single categorical X variable in a one-way MANOVA may be the single categorical Y variable in the corresponding DA, and the multiple Y variables in the one-way MANOVA may be the multiple X variables in the corresponding DA. Furthermore, some authors categorize DA as predictive and descriptive (Huberty, 1975, for example), and some as predictive and explanatory (Pedhazur, 1997, for example).

CA was initially developed in part to determine the relationship between a set of multiple continuous X variables and a set of multiple continuous Y variables. As practitioners became more knowledgeable about CA, some practitioners have become comfortable relative to the use of categorical variables as long as they are coded appropriately. Furthermore, other statistical techniques such as ANOVA, some MANOVAs and DAs, multiple regression analysis, and correlation may be looked upon as special cases of CA.

Some important statistics associated with MANOVA, DA, and CA are Wilks' Lambda (Λ), and standardized (β) and structure (r_s) coefficients.

Λ pertains to the relationship between the X and Y variable composites taking into account all the solutions or roots. β values pertain to the relationship between a variable in one set and a variable in or variate (a variable composite) of the other set controlling for the other variables in its own set. r_s values pertain to the relationship between a variable and the variate of its own set. These statistics are produced for each solution relative to both canonical and discriminant analyses.

Furthermore, these statistics are frequently utilized in practice, and reports relative to their stability (degree to which their values may be cross validated across samplings) exist.

However, the information that exists reflects conflicting points of view among notable statisticians –- for example, Tardif and Hardy (1995), Thompson (1991), Huberty (1975), and Barcikowski and Stevens (1975). This stability issue is largely dealt with in another paper that is being presented at this same meeting (Strand & Kossman, 2000).

Introduction to Specification Errors

Specification errors pertain to an inappropriate model with regard to the variables that are selected for study. The boundaries for what this means may range from narrow to broad (Pedhazur, 1997, p. 35) just as the meaning of "statistics" may range from narrow to broad. It is no wonder that beginners in statistics can be confused in their learnings! Relative to this paper a broad perspective was applied –- the specification errors include, but are not limited to, the following:

1. Variables whose distributions are "seriously" skewed: Among other requirements, a multivariate normal distribution is required for unambiguous interpretation of a variety of multivariate statistics as well as for the validity of important tests –- such as the test relative to Λ. A skewed distribution for any variable would contribute to a departure from the multivariate normal requirement.

2. Variables measured with at least considerable measurement error: The effects of measurement errors regarding multiple linear regression analysis (MLRA) have been particularly well studied –- for example see Cochran (1968), Guilford (1954), and Pedhazur (1997). Different types of measurement errors have been identified. Besides the effect of measurement error on the power of statistical tests, other likely even more serious consequences may occur. Predictable bias in the computation of R^2, b, and β occurs.

The biasing effects of measurement errors may be "complex" and may seriously alter the signs and absolute values of statistics such as β. Various "corrections" have been created and applied –- however each is accompanied by at least some negative consequence. A general recommendation with regard to all the issues associated with the consequences of measurement error is to attempt to measure variables with a high degree of validity –- often an "impossible" task, and to be sensitive to the often complex effects of the errors.

3. Failure to incorporate a nonlinear model when its use is justified: Among other effects, failure to utilize "powered terms" may decrease predictive and explained variance as well as result in insufficient representation of the relationships that do exist. These same points of view apply to the next type of specification error.

4. Failure to specify "interaction" variables when their use is justified.

5. Omitting relevant variables from the model: This type of specification error has been particularly well studied relative to MLRA and, relative to MLRA, is typically more serious than the next type of error listed. The most serious consequence of this type of error is likely the biased estimates of the regression coefficients. Even under the conditions that the estimates are not biased, the power of the relevant statistical tests is somewhat adversely affected by the omission of the relevant variables (Pedhazur, 1997).

6. Including irrelevant variables in the model (Rao, 1971; Mauro, 1990): While typically in MLRA this type of error is less serious than the previous error listed, the tests of significance concerning the relevant variables are typically with somewhat lower power than under the condition that the irrelevant variables are not included in the model.

7. Collinearity –- which translates in practice to relatively high correlations among within-set variables: Effects of collinearity, sometimes referred to as multicollinearity, is

understood well with regard to MLRA. Collinearity often has serious effects in MLRA –- most notably relative to the sign and magnitude of the regression coefficients. The effects may lead to misinterpretations at the least and in some cases result in "useless" regression coefficients and their associated statistical tests. Among other references, Mandel (1982) was of the opinion "Undoubtedly, the greatest source of difficulties in using least squares is the existence of 'collinearity' in many sets of data" (p. 15). This statement then applies to the DA and CA as well as MLRA.

Sparsity of Information Concerning Specification Errors As They Pertain to DA and CA

While more than a considerable amount of research has been conducted with regard to specification errors as they pertain to MLRA, an insufficient quantity of information exists relative to these errors as they pertain to DA and CA with the possible exceptions of studies concerning collinearity and the robustness of some multivariate tests.

The writer has not been able to find any study that pertains to the effects of not including powered or interaction terms when the data and/or literature suggest that one or both should be included. While the writer has not conducted an exhaustive study of this matter, in observing reports of hundreds of studies in which DA and/or CA were utilized the writer cannot recall any study in which powered or interaction terms were utilized –-even though in some cases they should have been. The writer suspects that seldom have these new variables been used in DA or CA.

Some studies do provide delimited insight into the effect of skewness relative to DA (for example –- Verboon & van der Lans, 1994; Randles, Broffitt, Ramberg, & Hogg, 1978; Broffitt, Clarke, & Lachenbruch, 1980). However, the sum of the results of all these studies that pertain to both DA and CA may be viewed as insufficient.

Even worse, with regard to DA and CA little has been studied concerning the issues of omission of relevant variables and inclusion of irrelevant variables as well as the effects of measurement errors. Somewhat limited information also exists regarding the effects of collinearity even though the studies that pertain to the stabilities of statistics such as standardized and structure coefficients in DA and CA sometimes also pertain to the collinearity issue.

The degree to which the specification errors have serious and complex effects is likely even greater in DA and CA than in MLRA.

Selected Suggestions Regarding How to Deal With Specification Errors as They Pertain to DA and CA

This section contains selected suggestions regarding how to deal with specification errors as they pertain to DA and CA. These are:

1. Seek updates regarding what the literature suggests concerning the effects of specification errors in DA and CA.

2. Attempt to measure variables with a high degree of validity. When the validity of measurement is not high, be appropriately sensitive about its consequences in interpreting statistics in DA and CA.

3. Be sensitive about the effects, often complex, of variables whose distributions are skewed and whose associated measurement is not notably valid.

4. Effectively consider and execute appropriately with regard to the use of powered and interaction terms in the DA and/or CA model.

5. Be sensitive to both omission of relevant variables and inclusion of irrelevant variables, especially the former, in both DA and CA.

6. Be most sensitive relative to collinearity –- not that it can be eliminated. When the collinearity is at least "considerable," be sensitive to its effects –- perhaps especially relative to standardized coefficients. Do not take lightly Mandel's (1982) statement –- "Undoubtedly, the greatest source of difficulties in using least squares is the existence of 'collinearity' in many sets of data" (p. 15).

References

Barcikowski, R. S., & Stevens, J. P. (1975). A Monte Carlo study of the stability of canonical correlations, canonical weights and canonical variate-variable correlations. *Multivariate Behavioral Research*, *10*(3), 353-364.

Broffitt, B., Clarke, W. R., & Lachenbruch, P. A. (1980). The effect of huberizing and trimming the Quadratic-discriminant function. *Communications in Statistics: Theory & Methods, 9*(1), 13-25.

Chateau, F. (1999). Structured discriminant analysis. *Communications in Statistics – Theory and Methods, 28*(2), 255-266.

Cochran, W. G. (1968). Errors of measurement in statistics. *Technometrics, 10*(4), 637-666.

Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1994). How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables. *Psychological Bulletin, 115*(3), 465-474.

Fung, W. K., & Gu, H. (1998). The second order approximation to sample influence curve in canonical correlation analysis. *Psychometrika, 63*(3), 263-269.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Huberty, C. J. (1975). The stability of three indices of relative variable contribution in discriminant analysis. *The Journal of Experimental Education, 44*(2), 59-64.

Huberty, C. J., & Wisenbaker, J. M. (1992). Variable importance in multivariate group comparisons. *Journal of Educational Statistics, 17*(1), 75-91.

Humphries-Wadsworth, T. M. (1998). *Features of published analyses of canonical results.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA. (ERIC Document Reproduction Service No. ED 418 125)

Mandel, J. (1982). Use of singular value decomposition in regression analysis. *The American Statistician, 36*(1), 15-24.

Mauro, R. (1990). Understanding L.O.V.E. (left out variables error): A method for estimating the effects of omitted variables. *Psychological Bulletin, 108*(2), 314-329.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Fort Worth: Harcourt Brace.

Randles, R. H., Broffitt, J. D., Ramberg, J. S., & Hogg, R. V. (1978). Generalized linear and quadratic discriminant functions using robust estimates. *Journal of the American Statistical Association, 73*(363), 564-568.

Rao, P. (1971). Some notes on misspecification in multiple regressions. *The American Statistician, 25*(5), 37-39.

Seo, T., Kanda, T., & Fujikoshi, Y. (1995). The effects of nonnormality of tests for dimensionality in canonical correlation and MANOVA models. *Journal of Multivariate Analysis, 52*(2), 325-337.

Strand, K. (1999). *The stability of standardized and structure coefficients: Further inquiry into their difference in each of canonical and discriminant analyses.* Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Strand, K., & Kossman, S. (2000). *Further inquiry into the stabilities of standardized and structure coefficients in canonical and discriminant analyses.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Tardif, B., & Hardy, J. (1995). Assessing the relative contribution of variables in canonical discriminant analysis. *Taxon, 44*(1), 69-76.

Thompson, B. (1991). Invariance of multivariate results: A Monte Carlo study of canonical function and structure coefficients. *The Journal of Experimental Education, 59*(4), 367-382.

Verboon, P., & van der Lans, I. A. (1994). Robust canonical discriminant analysis. *Psychometrika, 59*(4), 485-507.

Ziari, H. A., Leatham, D. J., & Ellinger, P. N. (1997). Development of statistical discriminant mathematical programming model via resampling estimation techniques. *American Journal of Agricultural Economics, 79*(4), 1352-1362.

Acknowledgement

Toni Waggoner made significant contributions to this paper.